Readability Measures and Automatic Text Simplification

Anonymous ACL submission

Abstract

Readability is a key concept in our era where textual information is abundant. Automatic text simplification (ATS) aims at making texts accessible to their target audience. Lately, there have been studies on the correlations between evaluation metrics in ATS and human judgment. However, the correlations between those two aspects and commonly available readability measures have not been the focus of as much attention. In this work, we investigate the place of readability measures in ATS by complementing the existing studies on evaluation metrics and human judgment. We first discuss the relationship between ATS and research in readability, then we report a study on correlations between readability measures and human judgment, and between readability measures and ATS evaluation metrics. We identify that LENS is the metric that correlates the most with readability measures. We find that for text simplification, lexical diversity is the type of feature that correlates the most with human judgment and evaluation metrics.

1 Introduction

The accessibility of written information is an important question: outside natural language processing, domains like medicine (Gu et al., 2024) or business (Huong Dau et al., 2024) have been studying the readability of the documents they produce (e.g. medical reports or information for patients, business reports for shareholders). Usually, those studies are performed using traditional readability formulas, like the famous Flesch Reading Ease (Flesch, 1948) or Dale-Chall (Dale and Chall, 1948) formulas. Recently, they have been acknowledging the reliability issues that come with those formulas (Alzaid et al., 2024). In natural language processing, Automatic text simplification (ATS) is a natural language processing (NLP) task that aims at transforming texts in order to make them more accessible, while preserving their meaning (Saggion, 2017). In ATS works, the goal is sometimes described as increasing the readability of a text. In this work, we investigate the place that readability occupies in the ATS landscape. We analyze the discourse on readability in ATS works by putting it in contrast with the lively field of automatic readability assessment (ARA), that aims at identifying the readability level of texts (Vajjala, 2022). While readability is regularly mentioned in current ATS works, ATS does not leverage ARA developments. Our contributions are the following: a discussion of ATS and ARA that identifies the bridges that remain to be made between the two fields; experiments with readability measures for ATS evaluation that fill a knowledge gap regarding correlations of evaluation practices and human judgment; insights for future developments for ATS evaluation and methods linked to readability.

2 Related Work

In this section, we discuss the fields of readability and text simplification that we introduce separately (Sections 2.1 and 2.2) before discussing how the two have interacted (Section 2.3).

2.1 Readability

Readability is a field of research that is considered to date back to the 1920's, with the first attempt to quantify the readability of English texts Lively and Pressey (1923). This first method relied on a list of word frequencies (Thorndike, 1921), where the more fequent the words of a text are, the more readable the text is condidered to be. François (2015) distinguishes several eras in text readability research, from Lively and Pressey (1923) to various paradigms of "AI readability". We synthesize this historical perspective below.

The early period consisted in identifying predictors and tune coefficients based on corpus-based observations and annotations from a given target audience. The most famous examples for English are Flesch Reading Ease (Flesch, 1948, FRE) and Flesch-Kincaid Grade Level (Kincaid et al., 1975, FKGL), which rely on word count and number of syllables per word.

The first approaches to measuring readability with NLP tools relied on linear regression on linguistic (i.e. syntactic and lexical) variables (Daoust et al., 1996), latent semantic analysis for textual coherence and cohesion (Foltz et al., 1998) and probabilities computed with language modeling (Si and Callan, 2001).

François (2015) concludes by noting an emerging trend at the time in ARA, that consists in relying on automatic feature extraction using neural networks. Ten years later, this has developed into a lively line of research (Vajjala, 2022). ARA has been explored with distributional text representations and with linguistic features. The distributional text representations follow the advancements of research in machine learning, notably with the development of transformers (Vaswani et al., 2017). Regarding linguistic features, the way to select and leverage them is still an open question. Nonetheless, research on this question is facilitate by the appearance of tools that can be used to compute an increasingly important number of features, for example for English (Kyle et al., 2021, 2018; Lu, 2010; Crossley et al., 2019) or French (Wilkens et al., 2022). Those tools produce raw analyses with hundreds of features, with no recommendations as to how to select and use them which is left up to the user. This has fueled research, notably with works that aim at combining those numeric representations with distributional representations (Deutsch et al., 2020; Lee et al., 2021; Wilkens et al., 2024).

The readability features depend heavily on the language that is under study. Indeed, the aforementioned tools rely on language-dependent resource such as reference corpora, vocabulary lists, or pretrained models (e.g. for POS-tagging or syntactic analysis).

2.2 Automatic Text Simplification

In this section, we briefly describe ATS to lay the ground for the discussion of how it integrates considerations about readability that comes in the next section (Section 2.3).

Methods. ATS has traditionally been performed at the sentence-level (Saggion, 2017). The goal was at first to make sentences simpler to handle as an input for other NLP systems such as syntactic parsers (Chandrasekar et al., 1996). It was only later explored as a means of simplifying texts to make them easier to understand by humans (Carroll et al., 1999). Those first methods were rule-based and targeted specific operations (Cardon and Bibal, 2023) such as removing appositive clauses or changing the voice of a sentence from passive to active. The recent developments of generative models has accelerated the shift of ATS research to document-level simplification (Sun et al., 2021), notably with multi-agent architectures (Mo and Hu, 2024; Fang et al., 2025) while sentence simplification is still being explored (Kew et al., 2023).

Evaluation. Evaluation of ATS is an open question. Traditional readability, mostly FKGL or adaptions of FRE for other languages are often reported, while it has been shown that they correlate poorly with the task (Tanprasert and Kauchak, 2021; Alva-Manchego et al., 2021). For sentence simplification, the most common metrics are BLEU (Papineni et al., 2002), SARI (Xu et al., 2016) – with an adaptation for document-level simplification D-SARI (Sun et al., 2021) – and BERTScore (Zhang et al., 2020). BLEU and BERTScore compare the output to one or more references, while (D-)SARI adds the input into the computation. Their correlation with the task is also unclear (Alva-Manchego et al., 2021; Sulem et al., 2018), although BLEU is often interpreted as an indicator of meaning preservation, SARI of simplicity, and BERTScore of meaning preservation and fluency.

Those three indicators are the three criteria that are used for human judgment to evaluate sentence simplification, typically on 5-point Likert scales. For document-level simplification, human evaluation is not stabilized. Cripwell et al. (2024) use the same criteria but using binary questions instead of Likert scales. Sun et al. (2021) ask judges to evaluate “overall simplicity" that they define as simplicity with other quality criteria such as ease of reading and meaning preservation. Vásquez-Rodríguez et al. (2023) ask judges to evaluate textual coherence. Agrawal and Carpuat (2024) evaluate meaning preservation by stuying human performance on reading comprehension tests.

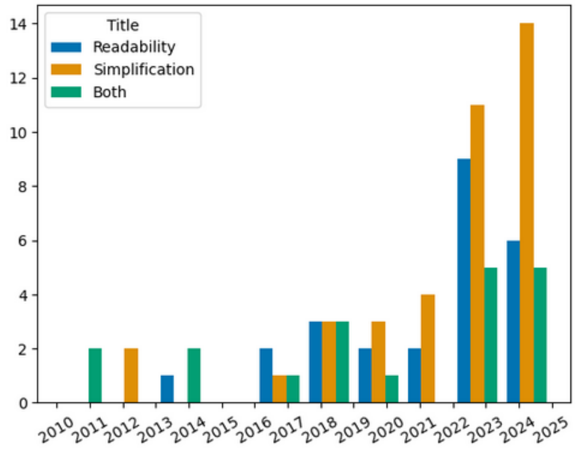


Figure 1: Number of papers from the ACL Anthology with “simplification" or “simplicity" in the title and “readability" in the abstract (“Simplification") or vice versa (“Readability") or both terms in the title (“Both").

2.3 Readability and Text Simplification

François and Bernhard (2014) (Vajjala and Meurers, 2014) To investigate the link between readability and text simplification, we extracted bibliographical data from the ACL Anthology, using the BibTeX Anthology with abstracts[1]. We extract papers with (i) the terms “readability" and “simplification" or “simplicity" in the title, (ii) “simplification" or “simplicity" in the title and “readability in the abstract and (iii) vice-versa. The number of results, plotted over time, is visible in Figure 1. We can see an increase of papers meeting those criterion over time. Table 1 displays information about the papers that have both “readability" and simplification in the title. Approximately a third of the papers (6 out of 16) concern English, three languages appear in two papers each (German, Portuguese – with Portuguese and Brazilian Portuguese – and Italian), and there is one paper for the following languages: Arabic, Chinese, Spanish, and Swedish.

8 out of 16 papers leverage readability for data analysis. All of those rely on features. Most of those works (6 out of 8) are resource papers and provide an analysis with readbility features to give information about the dataset (Battisti et al., 2020; Vajjala and Lučić, 2018; Yaneva et al., 2016; Štajner and Saggion, 2013; Dell’Orletta et al., 2011; Aluisio et al., 2010). Jingshen et al. (2024) rely on features for data selection instead, where readability features, in conjunction with similarity measures, are leveraged to mine sentence pairs to produce a parallel corpus for Chinese idiom simplification. De Martino (2023) investigates the link between eye-tracking data and readability features on Italian data. While it is a preliminary study, it suggests that eye-tracking is promising for evaluating the effect of simplification transformations.

Article	Lang.	Usage	Approach
Barayan et al. (2025)	EN	LLM Prompting	CEFR
Scholz and Wenzel (2025)	DE	Evaluation	Features
Jingshen et al. (2024)	ZH	Data analysis	Features
Paula and Camilo-Junior (2024)	PT-BR	Evaluation	Portuguese FRE
De Martino (2023)	IT	Data Analysis	Features and eye-tracking
Flores et al. (2023)	EN	Loss Component	Bounded FKGL
Engelmann et al. (2024)	EN	Evaluation	Formulas
Hazim et al. (2022)	AR	Visualization for manual simplification assistance	Lexical features
Battisti et al. (2020)	DE	Data analysis	Features
Maddela and Xu (2018)	EN	Lexical substitutes ranking	Lexical features
Vajjala and Lučić (2018)	EN	Data analysis	Features
Yaneva et al. (2016)	EN	Data analysis	Features
Grigonyte et al. (2014)	SV	Complexity identification	Lexical features
Štajner and Saggion (2013)	ES	Data analysis	Features / Formulas
Dell’Orletta et al. (2011)	IT	Data analysis	Features
Aluisio et al. (2010)	PT	Data analysis	Features / formulas

Table 1: Summary of papers of the ACL Anthology with both “readability" and “simplification" in the title. The table is sorted by descending year of publication.

The second most frequent use case is evaluation, with 3 papers. Scholz and Wenzel (2025) evaluate 18 readability features (syntactic, POS-based, semantic and fluency features) for English and German text simplification. Their findings is that some metrics are transferable (semantic, fluency), and that the behavior of statistical, POS-based and syntactic metrics seem to be strongly language-dependent. Paula and Camilo-Junior (2024) use a Portuguese adaption of FRE as an evaluation metric for ATS. (Engelmann et al., 2024) use the FRE and Dale-Chall formulas to perform pairwise comparisons in an Elo-like ranking system. They compare it to human judgments and GPT 3.5 performance. They find that Dale-Chall has the highest correlation to human judgment, above GPT 3.5, while FRE obtains the lowest correlations.

3 papers use lexical complexity features for lexical simplification (North et al., 2025). Hazim et al.

[1]Available at `https://aclanthology.org/anthology+abstracts.bib.gz`.

(2022) introduce a system that highlights complex words in a text editor to help humans manually simplify texts. Maddela and Xu (2018) use lexical features to rank candidates for substitution in a neural lexical simplification system. (Grigonyte et al., 2014) rely on features to perform complex word identification.

Finally, Flores et al. (2023) use a bounded FKGL (ranging from 4 to 20, based on empirical observations) as a component of their loss in a neural model for text simplification. (Maddela and Alva-Manchego, 2025) prompt LLMs for document-level simplification by including CEFR levels in the prompt, as was also done by Imperial and Tayyar Madabushi (2023). Using CEFR as a proxy for readability is a trend that was initiated with the release of the CEFR-SP dataset (Arase et al., 2022).

In conclusion, we observe that different approaches to readability (features, formulas, eye-tracking, CEFR levels) are explored in ATS works. The two approaches that are widely present in ATS are traditional formulas, which have consistently been used as an evaluation metric, and readability features, that have been used to give information about datasets. In this work, we explore how features correlate with human judgment on the simplification task.

3 Studying Correlations between Readability Measures and ATS Metrics

3.1 Data

In order to study how readability features correlate with the evaluation protocols in ATS, we rely on data that is labeled with human judgment and on which automatic metrics can be computed. Two studies provide this kind of data, at the sentence level (Alva-Manchego et al., 2021) and at the document level (Maddela and Alva-Manchego, 2025). Both studies aim at studying the link between automatic metrics and human judgment. To this, we add observations on the link between readability measures and human judgment, and on the link between readability measures and automatic metrics. We describe the datasets below.

SimplicityDA. For the sentence-level study, we use Simplicity-DA (Alva-Manchego et al., 2021)[2]. It is a set of 600 sentence simplification system outputs in English, each one annotated by 15 crowd-workers along the three common human judgment criteria in ATS: fluency, simplicity and meaning preservation. The dataset also includes automatic scores for each sentence: BLEU, SARI, BERTScore and SAMSA.

For the document-level study, we use D-Wikipedia (Sun et al., 2021). D-Wikipedia is a corpus of aligned paragraph pairs that come from the English Wikipedia for the complex side and Simple English Wikipedia for the simple side. Maddela and Alva-Manchego (2025) released a subset of 100 paragraph pairs from D-Wikipedia, each with 4 automatic simplifications, resulting in 500 paragraph pairs. Those 500 pairs were rated by three human judges on fluency, simplicity and meaning preservation. We compute the automatic metrics values with the code provided with the dataset[3]. Those automatic metrics are BLEU, SARI, D-SARI, BERTScore and LENS. Maddela and Alva-Manchego (2025) also introduce adaptations of SARI, LENS and BERTScore (respectively Agg-SARI, Agg-LENS and Agg-BERTScore) to the document-level simplification task by aggregating scores computed at the sentence-level.

Tool	Type	Nb	List of features
TAALES	Lexical Sophistication	485	Link
TAACO	Cohesion	168	Link
TAASSC	Syntactic Sophistication	355	Link
TAALED	Lexical Diversity	38	Link

Table 2: Summary of the tools used for readability features in this study, with links to the lists of features and their description.

3.2 Readability Measures

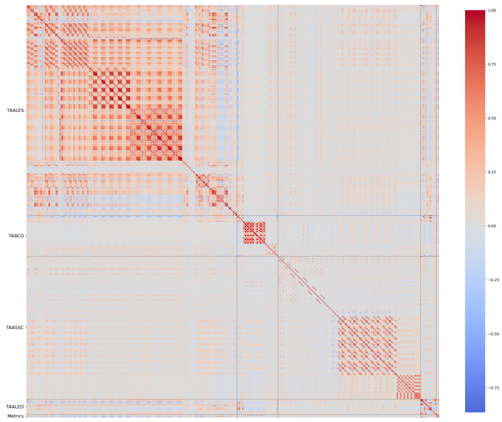
Readability Features. As discussed in Section 2, readability is now mostly explored with two types of text representations: distributional embeddings and textual features. As distributional embeddings are leveraged for ATS methods and evaluation, we focus on textual features. To compute those features, we use what we find to be the most extensive suite of tools for computing readability measures: TAALED (Kyle et al., 2021), TAALES (Kyle et al., 2018), TAASSC (Lu, 2010) and TAACO (Crossley et al., 2019)[4]. Table 2 details the characteristics of what each tool is used for.

Readability Metrics. We also compute the following series of traditional readability metrics us-

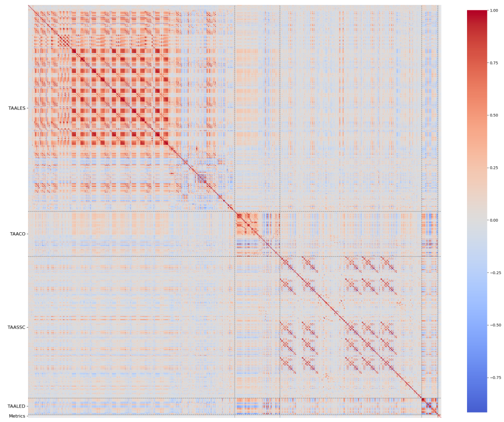
[2]https://github.com/feralvam/metaeval-simplification

[3]https://github.com/cardiffnlp/document-simplification

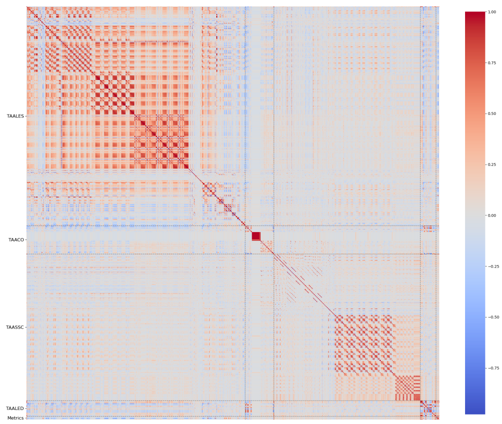
[4]All available at https://www.linguisticanalysistools.org/



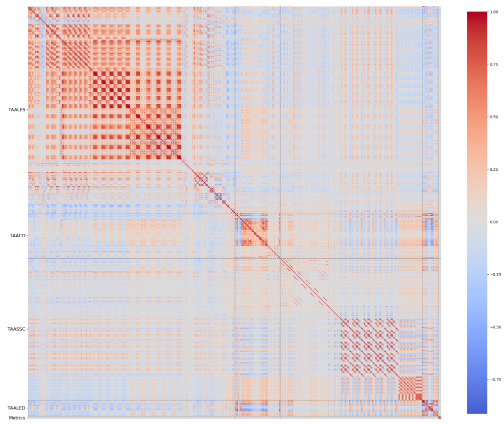
(a) Difference between original and simple, SimplicityDA (sentence-level).



(b) Difference between original and simple, D-Wikipedia (document-level).



(c) Simple side of SimplicityDA (sentence-level).



(d) Simple side of D-Wikipedia (document-level).

Figure 2: Pearson correlation matrices of readability measures and metrics. Dashed lines indicate the boundaries of feature groups (from top to bottom, and the same from left to right: TAALES, TAACO, TAASSC, TAALED, and Metrics).

ing the `textstat` Python library: Flesch Reading Ease (Flesch, 1948), Dale-Chall (Dale and Chall, 1948), Gunning-Fog (Gunning, 1952), Linsear Write (O'hayre, 1966), ARI (Smith and Senter, 1967), SMOG (Mc Laughlin, 1969), Flesch-Kincaid Grade Level (Kincaid et al., 1975), and Coleman-Liau (Coleman and Liau, 1975).

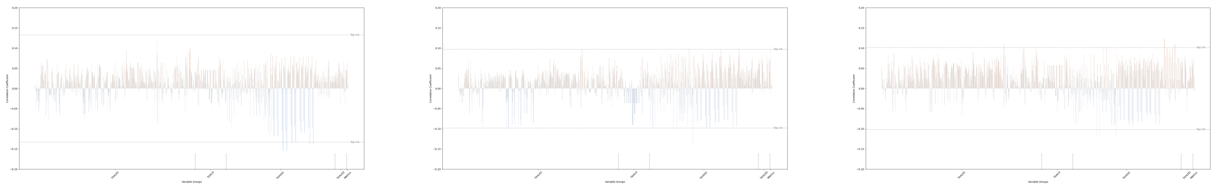
4 Experiments

4.1 Readability Measures

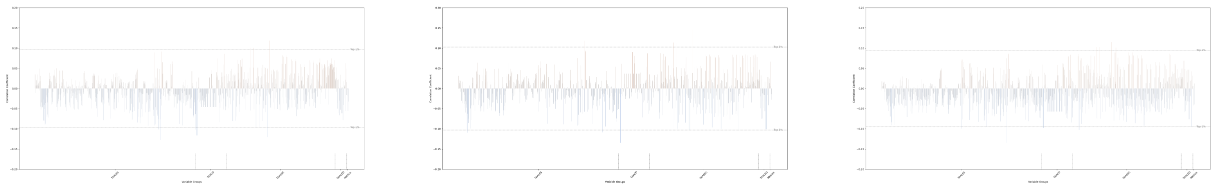
First, we compute the correlations between the readability measures (metrics and features) themselves. Figures 2a and 2c show the correlation matrices computed on the SimplicityDA dataset (at the sentence level), respectively on the difference between the simplified and original sentences, and on the simplifications. Figures 2b and 2d show the correlation matrices computed on the D-Wikipedia dataset, respectively on the difference between the simplified and original sentences, and on the simplifications. We make three observations: (i) the measures mostly correlate with other measures of the same type, (ii) measures computed at the document-level show higher absolue values and (iii) measures computed on the difference between original texts and simplifications exhibit lower absolute values.

4.2 Measures and Human Judgment

To compare readability measures (the features with the four readability tools, and the readability metrics) and human judgment, we compute them all on both datasets: SimplicityDA for the sentence-level (100 original sentences and 600 simplifications including 100 human-written ones) and D-Wikipedia for the document-level (100 original paragraphs and 500 simplifications including 100 human-written ones). For each dataset we compute the measures on both sides (original and simplified) separately. We compute the correlations with

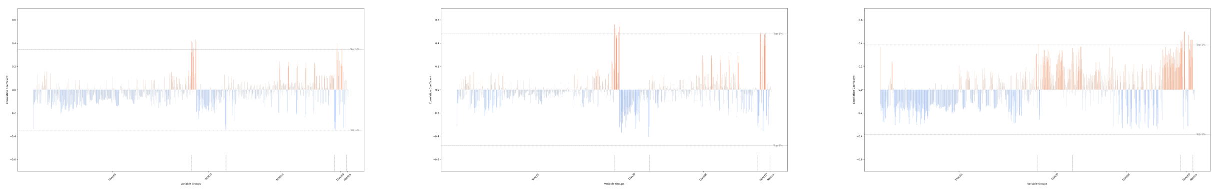


(a) Simplicity (b) Fluency (c) Meaning

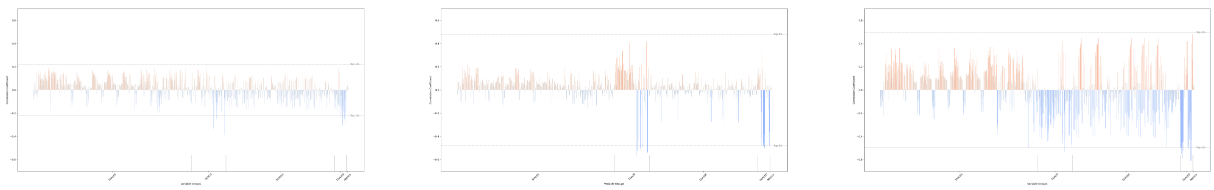


(d) Simplicity (e) Fluency (f) Meaning

Figure 3: Correlations between readability measures and human judgment criteria on the SimplicityDA dataset (sentence-level). The first row shows the correlations with the simplifications, while the second row shows the correlations with the difference between the original and simplified texts. X-axis represents the readability measures, by group (from left to right TAALES, TAACO, TAASSC, TAALED, Metrics) while Y-axis indicates the correlation values on a scale from -0.2 to 0.2. Horizontal lines represent the threshold of the top 1% absolute values. Color vividness indicates the absolute value of the correlation.



(a) Simplicity (b) Fluency (c) Meaning



(d) Simplicity (e) Fluency (f) Meaning

Figure 4: Correlations between readability measures and human judgment criteria on the D-Wikipedia dataset (document-level). The first row shows the correlations with the simplifications, while the second row shows the correlations with the difference between the original and simplified texts. X-axis represents the readability measures, by group (from left to right TAALES, TAACO, TAASSC, TAALED, Metrics) while Y-axis indicates the correlation values on a scale from -0.7 to 0.7. Horizontal lines represent the threshold of the top 1% absolute values. Color vividness indicates the absolute value of the correlation.

human judgment in two ways: (i) on the measures obtained on the simplifications only, and (ii) on the difference between the measures obtained on the original texts and the ones obtained on the simplifications. The first case focuses on simplicity, the second case focuses on simplification, by including a comparison with the original text.

For both datasets, we report the correlations on the three criteria for human judgment: simplicity, fluency and meaning preservation.

4.3 Measures and Automatic Metrics

To study the correlations between readability measures and automatic ATS metrics, we proceed in the same way as for the correlations between readability measures and human judgment. We report scores on the following automatic metrics: BLEU, SARI, BERTScore, SAMSA for simplicityDA, and BLEU, SARI, D-SARI, BERTScore, LENS, Agg-SARI, Agg-LENS and Agg-BERTScore for D-Wikipedia.

For the metrics that require references, for Simplicity-DA we use all the references that are provided, i.e. for each original sentence 10 references from ASSET (Alva-Manchego et al., 2020), 1 from TurkCorpus (Xu et al., 2016) and 1 from HSplit (Sulem et al., 2018). For D-Wikipedia, we use the one reference simplification that is provided for each original text.

5 Results

5.1 Measures and Human Judgment

We report the correlations between readability measures and human judgment at Figure 3 for the SimplicityDA dataset, and Figure 4 for D-Wikipedia.

For SimplicityDA, the correlations are very low: the highest absolute value across all variables and criteria is at obtained with the variable `news_av_delta_p_const_cue` (TAASSC), with a correlation coefficient of -0.16.

Regarding the D-Wikipedia dataset, we can see that the readability measures correlate better with the human judgment than on SimplicityDA. Simplicity is the criterion that has the lowest top 1% threshold is simplicity, with a threshold of 0.35 when computed on simplifications only, and at 0.22 on the difference between original texts and simplifications. For fluency, those values are both at 0.48, and for meaning respectively at 0.38 and 0.50. The top variables for simplicity are different kinds of type/token ratios (from TAACO and TAALED), i.e. on lemmas, content words and nouns, for both ways of computing the values.

5.2 Measures and Automatic Metrics

We report the correlations between readability measures and automatic metrics at Figure 5 for the SimplicityDA dataset, and Figure 6 for D-Wikipedia.

For SimplicityDA, SARI and BERTScore have the highest correlation values: the threshold for the top 1% of absolute values is at 0.41 for both (computed on the difference between original and simplified texts). SAMSA exhibits the lowest correlation, with a threshold at 0.25 on simplifications and 0.18 on the difference. BLEU has a threshold at 0.28 for both computations. While the top metrics vary according to the setting (metric and computation), they consistently come from TAALES and TAALED, indicating that for this set of observations, lexical features are the most relevant ones.

Regarding D-Wikipedia, the correlations are generally higher. The highest ones are obtained with LENS: 0.50 on simplifications and 0.51 on the difference between original texts and simplifications. A notable observation is the difference between SARI and BERTScore and their adaptations: on simplifications, SARI obtains 0.20 and Agg-SARI 0.29 , BERTScore obtains 0.10 and Agg-BERTScore 0.30. On the difference, those numbers are at 0.18 and 0.20 for SARI, and at 0.10 and 0.31 for BERTScore. This increase is not observed with LENS, as Agg-LENS obtains 0.43 (vs 0.50 for LENS) on simplifications. For all LENS and Agg-LENS results, the top features are all related to lexical diversity with different kinds of type/token ratios (lemma, content words, bigram, nouns).

6 Discussion

In this section, we summarize the main findings of our study and discuss their implications. Readability measures are more adapted to work at the document-level than at the sentence-level. We make those observations both on correlations with human judgments and automatic metrics.

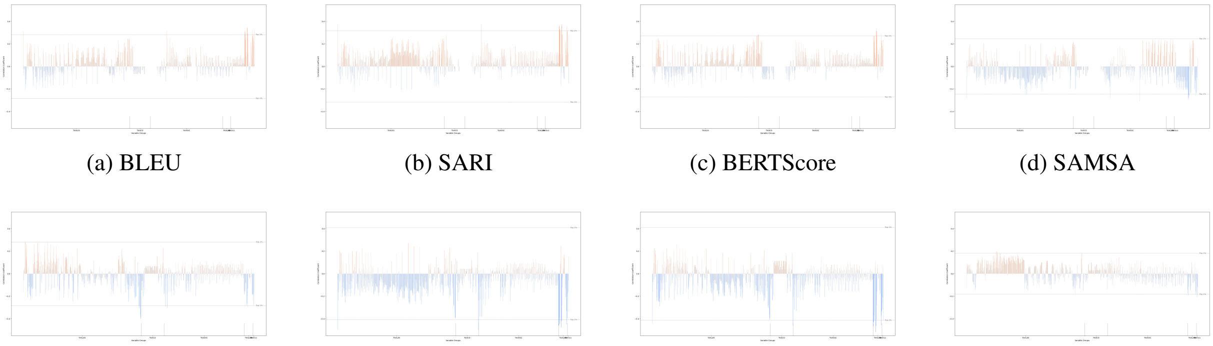
Most automatic metrics do not correlate with readability measures. LENS is a notable exception, with correlations that can go up to 0.61 (lemma type/token ratio) for the highest value. The aggregation method proposed by Maddela and Alva-Manchego (2025) substantially increases the correlations between readability measures and the two metrics SARI and BERTScore. Traditional formulas consistently have low correlation values.

Regarding the kind of variables that display the higher correlations, we consistently find variables related to lexical diversity, and more precisely various kinds of computing the type/token ratio. This suggests that focusing on ways of measuring and integrating lexical diversity in the works on ATS systems may be a promising direction.

Regarding future directions, on top of judgments from identified groups, further research with eye-tracking analyses may help inform on what aspects should be the focus of evaluation.

7 Conclusion

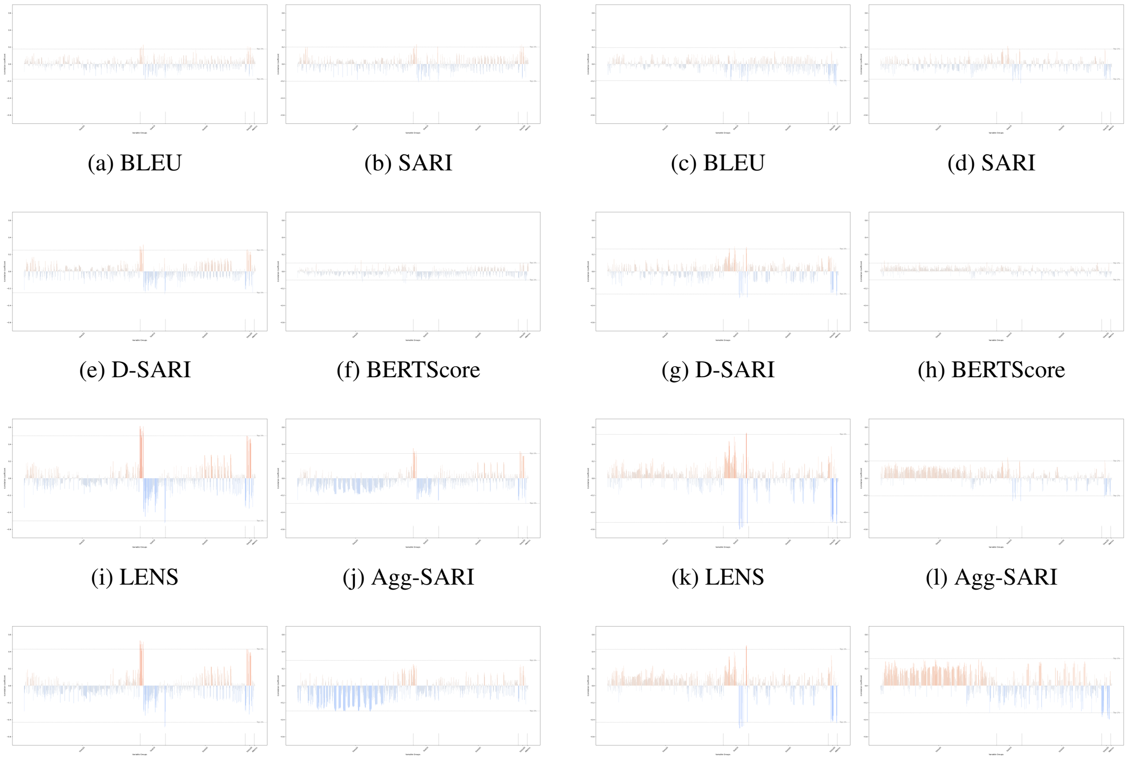
In this study, we explored the correlations between readability measures and human judgment, and between readability measures and automatic metrics. We found that the correlations are in the same range as the ones displayed when studying automatic metrics and human judgment. We found that lexical diversity features seem to be the type of features



(a) BLEU (b) SARI (c) BERTScore (d) SAMSA

(e) BLEU (f) SARI (g) BERTScore (h) SAMSA

Figure 5: Pearson correlations between readability measures and automatic ATS metrics, on SimplicityDA. The readability values are computed on the simplifications (first row) and on the difference between the original texts and the corresponding simplifications (second row). X-axis represents the readability measures, by group (from left to right TAALES, TAACO, TAASSC, TAALED, Metrics) while Y-axis indicates the correlation values on a scale from -0.55 to 0.55. Horizontal lines represent the threshold of the top 1% absolute values. Color vividness indicates the absolute value of the correlation.



(a) BLEU (b) SARI (c) BLEU (d) SARI

(e) D-SARI (f) BERTScore (g) D-SARI (h) BERTScore

(i) LENS (j) Agg-SARI (k) LENS (l) Agg-SARI

(m) Agg-LENS (n) Agg-BERTScore (o) Agg-LENS (p) Agg-BERTScore

Figure 6: Pearson correlations between readability measures and automatic ATS metrics, on D-Wikipedia. The readability values are computed on the simplifications (columns 1-2) and on the difference between the original texts and the corresponding simplifications (columns 3-4). X-axis represents the readability measures, by group (from left to right TAALES, TAACO, TAASSC, TAALED, Metrics) while Y-axis indicates the correlation values on a scale from -0.7 to 0.7. Horizontal lines represent the threshold of the top 1% absolute values. Color vividness indicates the absolute value of the correlation.

that is the most correlated to the simplification task. With this work, combined on the observations made on automatic metrics and human judgment on the same data, we have an idea of the interactions between automatic metrics, human judgment, and readability measures.

8 Limitations

As discussed in Section 2, readability measures are rather language-dependent. We conducted this study on English because data with human judgments, both at the sentence-level and at the document-level, are readily available.

Also, this study involves only two datasets. It is unclear whether our observations would generalize to other datasets. Quality human-labeled datasets are scarce, this limitation is one of the domain.

References

Sweta Agrawal and Marine Carpuat. 2024. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Mohammad Alzaid, Faisal R Ali, and Emma Stapleton. 2024. Limitations of readability assessment tools. *European Archives of Oto-Rhino-Laryngology*, 281(9):5021–5022.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A corpus for automatic readability assessment and text simplification of German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France. European Language Resources Association.

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–130, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270, Bergen, Norway. Association for Computational Linguistics.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2024. Evaluating document simplification: On the importance of separately assessing simplicity and meaning preservation. *LREC-COLING 2024*, page 1.

Scott A Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, 51:14–27.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

François Daoust, Léo Laroche, and Lise Ouellet. 1996. Sato-calibrage: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique*, 25(1):205–234.

Maria De Martino. 2023. Processing effort during reading texts in young adults: Text simplification, readability assessment and preliminary eye-tracking data. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 179–184, Venice, Italy. CEUR Workshop Proceedings.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Björn Engelmann, Christin Katharina Kreutz, Fabian Haak, and Philipp Schaer. 2024. ARTS: Assessing readability & text simplicity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14925–14942, Miami, Florida, USA. Association for Computational Linguistics.

Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative document simplification using multi-agent systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912, Abu Dhabi, UAE. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Lorenzo Jaime Flores, Heyuan Huang, Kejian Shi, Sophie Chheang, and Arman Cohan. 2023. Medical text simplification: Optimizing for readability with unlikelihood training and reranked beam search decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4859–4873, Singapore. Association for Computational Linguistics.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Thomas François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, XX(2):79–97.

Thomas François and Delphine Bernhard. 2014. When text readability meets automatic text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):89–96.

Gintarė Grigonyte, Maria Kvist, Sumithra Velupillai, and Mats Wirén. 2014. Improving readability of Swedish electronic health records through lexical simplification: First results. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 74–83, Gothenburg, Sweden. Association for Computational Linguistics.

Joey Z Gu, Grayson L Baird, Antonio Escamilla Guevara, Young-Jin Sohn, Melis Lydston, Christopher Doyle, Sarah EA Tevis, and Randy C Miles. 2024. A systematic review and meta-analysis of english language online patient education materials in breast cancer: Is readability the only story? *The Breast*, page 103722.

Robert Gunning. 1952. The technique of clear writing. *McGraw-Hill*.

Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2022. Arabic word-level readability visualization for assisted text simplification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–249, Abu Dhabi, UAE. Association for Computational Linguistics.

Nam Huong Dau, Duy Van Nguyen, and Hai Thi Thanh Diem. 2024. Annual report readability and firms' investment decisions. *Cogent Economics & Finance*, 12(1):2296230.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

Zhang Jingshen, Chen Xinglu, Qiu Xinying, Wang Zhimin, and Feng Wenhe. 2024. Readability-guided idiom-aware sentence simplification (RISS) for Chinese. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1183–1200, Taiyuan, China. Chinese Information Processing Society of China.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

J Peter Kincaid, RP Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (automated reliability index, fog count and flesch reading ease formula) for navy enlisted personnel (research branch report 8-75). memphis, tn: Naval air station; 1975. *Naval Technical Training, US Naval Air Station: Millington, TN*.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (taales): Version 2.0. *Behavior research methods*, 50:1030–1046.

Kristopher Kyle, Scott A Crossley, and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18(2):154–170.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bertha A Lively and Sidney L Pressey. 1923. A method for measuring the vocabulary burden of textbooks. *Educational administration and supervision*, 9(7):389–398.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Mounica Maddela and Fernando Alva-Manchego. 2025. Adapting sentence-level automatic metrics for document-level simplification evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6444–6459, Albuquerque, New Mexico. Association for Computational Linguistics.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Kaijie Mo and Renfen Hu. 2024. ExpertEase: A multi-agent framework for grade-specific document simplification with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099, Miami, Florida, USA. Association for Computational Linguistics.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2025. Deep learning approaches to lexical simplification: A survey. *Journal of Intelligent Information Systems*, 63(1):111–134.

John O'hayre. 1966. *Gobbledygook has gotta go*. US Department of the Interior, Bureau of Land Management.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Antonio Flavio Paula and Celso Camilo-Junior. 2024. Evaluating the simplification of Brazilian legal rulings in LLMs using readability scores as a target. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 117–125, Miami, Florida, USA. Association for Computational Linguistics.

Horacio Saggion. 2017. *Automatic Text Simplification*. Morgan & Claypool Publishers.

Karen Scholz and Markus Wenzel. 2025. Evaluating readability metrics for German medical text simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6049–6062, Abu Dhabi, UAE. Association for Computational Linguistics.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576.

E.A. Smith and R.J. Senter. 1967. *Automated Readability Index*. AMRL-TR. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command.

Sanja Štajner and Horacio Saggion. 2013. Readability indices for automatic evaluation of text simplification systems: A feasibility study for Spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 374–382, Nagoya, Japan. Asian Federation of Natural Language Processing.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Edward L Thorndike. 1921. Word knowledge in the elementary school. *Teachers College Record*, 22(4):1–27.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2):194–222.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. Document-level text simplification with coherence evaluation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P. Yancey, and Thomas François. 2022. FABRA: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233, Marseille, France. European Language Resources Association.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian's, Malta. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2016. Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 293–299, Portorož, Slovenia. European Language Resources Association (ELRA).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.